REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
11/06/2009	FINAL REPORT	03/30/2007-11/6/2009

4. TITLE AND SUBTITLE

Adaptive Sensing and Fusion of Multi-Sensor Data and Historical Information

5a. CONTRACT NUMBER

N00014-07-C-0357

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. Lawrence Carin,

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Signal Innovations Group, Inc.
1009 Slater Rd., Ste. 200
Durham, NC 27703

8. PERFORMING ORGANIZATION REPORT NUMBER

SIG.ONR.037.FINAL

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
875 N. Randolph St. Ste. 1425
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

A

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Context plays an important role when performing underwater classification, and in this report we examine context from two perspectives. First, the classification of items within a single task is placed within the context of distinct concurrent or previous classification tasks (multiple distinct data collections). This is referred to as multi-task learning (MTL), and is implemented here in a statistical manner, using a simplified form of the Dirichlet process. In addition, when performing many classification tasks one has simultaneous access to all unlabeled data that must be classified, and therefore there is an opportunity to place the classification of any one feature vector within the context of all unlabeled feature vectors; this is referred to as semi-supervised learning. In this report we integrate MTL and semi-supervised learning into a single framework

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Samantha Venters
U	U	U	U	2	19b. TELEPHONE NUMBER *(Include area code)* 919-323-3453

FINAL REPORT

Adaptive Sensing and Fusion of Multi-Sensor Data and Historical Information

Lawrence Carin
Signal Innovations Group, Inc.
1009 Slater Rd. Suite 200
Durham, NC 27703
phone: (919) 475-2151 fax: 919-660-5293 email: lcarin@ee.duke.edu

Award Number N00014-07-C-0357

LONG-TERM GOALS

This project has focused on development of active learning and semi-supervised learning algorithms for underwater sensing. The research is being executed in collaboration with NSWC Panama City, which serves as a good source of data for algorithm testing, and a transition point for the algorithms. A long-term goal is to transition the active-learning technology to the NSWC software suite.

OBJECTIVES

Context plays an important role when performing underwater classification, and in this report we examine context from two perspectives. First, the classification of items within a single task is placed within the context of distinct concurrent or previous classification tasks (multiple distinct data collections). This is referred to as multi-task learning (MTL), and is implemented here in a statistical manner, using a simplified form of the Dirichlet process. In addition, when performing many classification tasks one has simultaneous access to all unlabeled data that must be classified, and therefore there is an opportunity to place the classification of any one feature vector within the context of all unlabeled feature vectors; this is referred to as semi-supervised learning. In this report we integrate MTL and semi-supervised learning into a single framework, thereby exploiting two forms of contextual information.

A key new objective of the research is to adapt the features to the environment. For this purpose we have introduced the Beta Process, the development and application of which have been an important component of the research executed over the last year, and reported here.

APPROACH

The Dirichlet process (DP) [1], [2], [3], denoted as $\mathcal{DP}(\alpha G_0)$, is a prior used in non-parametric Bayesian models, for the purpose of clustering data. It is a distribution over probability measure, i.e., each draw G from a Dirichlet process is itself a distribution. The base measure G_0 is the prior mean of the DP and the concentration parameter α, acting as the inverse variance, controls how much the draw G from a DP is allowed to deviated from the base measure G_0. The larger α is, the smaller the variance is, and G will concentrate more of its mass around the mean G_0. In the limit as $\alpha \rightarrow \infty$, G goes to G_0 weakly or pointwise. However, we should note that we cannot say that G goes to G_0 in the limit as $\alpha \rightarrow \infty$ since draws from a DP will be discrete distributions with probability one, even if the base measure G_0 is continuous. In the limits as $\alpha \rightarrow 0$, G takes random discrete values.

Let $\{\theta_1, \theta_2, \ldots, \theta_n\}$ be a sequence of independent draws from a prior G, with G itself a sample from a $\mathcal{DP}(\alpha G_0)$. The mathematical representation of the DP model is:

$$\begin{aligned} \theta_i &\sim G \\ G &\sim DP(\alpha G_0) \end{aligned} \tag{1}$$

Marginalizing out G, the conditional distribution θ_{N+1} given the other N observations

$\{\theta_1, \theta_2, \ldots, \theta_N\}$ is

$$p(\theta_{N+1}|\theta_1, \theta_2, \ldots, \theta_N, \alpha, G_0) = \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\theta_n} + \frac{\alpha}{\alpha + N} G_0 \tag{2}$$

where δ_{θ_n} is a point mass located at θ_n.

Note that the draw G from a DP is discrete, and consequently multiple θ_i's may take the same value simultaneously. Let θ_k^*, $k = 1, 2, \ldots, K$, denote K distinct values among $\theta_1, \theta_2, \ldots, \theta_N$ and n_k be the number of repeats of θ_k^*. The conditional distribution of (2) can be equivalently written as:

$$p(\theta_{N+1}|\theta_1, \theta_2, \ldots, \theta_N, \alpha, G_0) = \frac{n_k}{\alpha + N} \sum_{k=1}^{K} \delta_{\theta_k^*} + \frac{\alpha}{\alpha + N} G_0 \tag{3}$$

From (3), we notice that the probability of θ_N taking the value of θ_k^* is proportional to n_k. The larger the n_k is, the more probable θ_N will take the value θ_k^*. This phenomenon can be called a clustering property since a new observation tends to join the group with a larger number of samples.

As discussed above, draws from a DP are discrete distributions. From the stick-breaking construction [4] this points is made more explicit by providing a constructive form of a draw from a DP. It is simply given as follows:

$$\begin{aligned} G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) \\ \beta_k &\sim \mathrm{Beta}(1, \alpha) \\ \theta_k^* &\sim G_0 \end{aligned} \tag{4}$$

Here $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{\infty} \pi_k = 1$. It is clear from the construction form of G that draws from a DP are discrete, composed of an infinite weighted sum of point masses. The construction of π can be understood as follows. Starting with a stick of length 1, we break it at $\beta_1 \sim \mathrm{Beta}(1, \alpha)$ and assign π_1 to be the length of stick we just broke off. Recursively breaking the remaining portion at $\beta_2, \beta_2, \ldots$, we get the length π_2, π_3 and so forth. Since the lengths π_k decrease stochastically with k, the summation in (4) may be truncated with N terms $G = \sum_{k=1}^{N} \pi_k \delta_{\theta_k^*}$, yielding an N level truncated approximation to a draw G from the DP [5]. In [5] is gaven a bound for the error introduced by the truncation in the DP.

Assuming we have a set of data $\{x_1, x_2, \ldots, x_N\}$ with associated hidden parameters $\{\theta_1, \theta_2, \ldots, \theta_N\}$, each θ_n is drawn independently and identically from G, while each x_n has distribution $F(\theta_n)$. Since G is discrete and multiple θ_n's may take the same value θ_k^*, datapoints x_n's associated with the same value θ_k^* belong to one cluster. Such a kind of model of data can be viewed as a mixture model with countable infinite components. Let z_i be a cluster assignment variable, which takes on value k with probability π_k.

The generative infinite mixture model can be expressed as

$$\begin{aligned} x_i|z_i, \{\theta_k^*\} &\sim F(\theta_{z_i}^*) \\ z_i|\pi &\sim Mult(\pi) \\ \pi|\alpha &\sim stick(\alpha) \\ \theta_k^* &\sim G_0 \end{aligned} \tag{5}$$

where $stick(\alpha)$ is stick-breaking process with paramter α. Different from the finite mixture model, which uses a fixed number of clusters to model the data, the number of distinct values of θ_n with a DP prior is driven by data as well as the concentration parameter α. In our work instead of setting a fixed value for α, a Gamma hyper-prior over α is employed, which yields greater model flexibility.

The beta process (BP) was first introduced by Hjort [6] for applications in survival analysis. A beta process $\mathcal{BP}(cB_0)$ is an independent increments or Lévy process with concentration parameter c and base measure B_0. Let Ω be a measurable space and $\mathcal{B}$ its σ-algebra. For all disjoint, infinitesimal partition $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_K$ of Ω, the Beta process is generated as

$$H(\mathcal{B}_k) \sim Beta(cB_0(\mathcal{B}_k), c(1 - B_0(\mathcal{B}_k))) \tag{6}$$

A draw from a Beta process ($\mathcal{BP}$) can be constructed as follows

$$\begin{aligned} B &= \sum_k \pi_k \delta_{\omega_k} \\ \omega_k &\sim B_0 \\ \pi_k &\sim Beta(cB_0(\omega_k), c(1 - B_0(\omega_k))) \end{aligned} \tag{7}$$

where δ_{ω_k} is a unit mass concentrated at ω_k ($\omega_k \in \Omega$). Note from (7), $\sum_{k=1}^{\infty} \pi_k \neq 1$, therefore B can not be treated as a probability mass function.

The two-parameter Beta process $\mathcal{BP}(cB_0)$ can be extended to a three parameter Beta process $\mathcal{BP}(a, b, B_0)$ [7], which is specifically as

$$H(\mathcal{B}_k) \sim Beta(aB_0(\mathcal{B}_k), b(1 - B_0(\mathcal{B}_k))) \tag{8}$$

Another interesting process closely related to Beta process is Bernoulli process [8], denoted as $X \sim BeP(B)$, where B is a measure on Ω. If B is continuous, X is a Poisson process with intensity B and can be constructed as

$$X = \sum_{i=1}^{N} \delta_{\omega_i} \tag{9}$$

where $N \sim Poi(B(\Omega))$ and ω_i are independent draws distribution from $B/B(\Omega)$. In the case for which B is discrete and let $B = \sum_k \pi_k \delta_{\omega_k}$, X has following construction form,

$$\begin{aligned} X &= \sum_k z_k \delta_{\omega_k} \\ z_k &\sim Bernoulli(\pi_k) \end{aligned} \tag{10}$$

X is then a set of elements which only take value $\{0, 1\}$ at different locations ω_k. For our application, Ω can be thought as a space of potential features and X as an

observation (or a datum) which possesses a part of features. The possession of features are different for different data and determined by z_k.

Let $\{X_1, X_2, \ldots, X_n\}$ be n independent draws of $BeP(B)$ from discrete B and B is a draw from $\mathcal{BP}(cB_0)$. Marginalizing out B, the predictive conditional probability of a new draw X_{n+1} given the previous draws $\{X_1, X_2, \ldots, X_n\}$ is

$$\begin{aligned} X_{n+1}|X_1, X_2, \ldots, X_n &\sim BeP(\frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} X_i) \\ &= BeP(\frac{c}{c+n}B_0 + \sum_{k=1}^{K}\frac{m_{n,k}}{c+n}\delta_{w_k}) \end{aligned} \tag{11}$$

where $m_{n,k}$ is the number of n data having possessed feature w_k.

Let $c = 1$, $\gamma = B_0(\Omega)$ and $\{X_1, X_2, \ldots, X_{n+1}\}$ be generated in sequence. This generating process is then reduced to Indian buffet process $IBP(\gamma, n)$[9],

- For the first observation (custom) X_1, the number of features possessed (or the number of dishes the customer tastes) is $Poi(B_0(\Omega))$ or $Poi(\gamma)$;
- For subsequent observations (customs) $X_{i+1}, i = 1, 2, \ldots, n$, the probability of selecting previous features (or old dishes) ω_k is $\frac{m_{i,k}}{1+i}$, where $\frac{m_{i,k}}{1+i}$ is the number of previous i observations (customers) selecting feature (dish) ω_k; the number of new features (or dishes) X_i will select is $Poi(B_0(\Omega)/(i+1)) = Poi(\gamma/(i+1))$.

As mentioned in [9], the Indian buffet process is the limiting case of a finite feature model as K the number of potential features tends to infinity. The finite feature model provides a full conjugacy and will allow for variational inference to be performed on the multi-task feature selection. The finite latent feature model may be defined as

$$\begin{aligned} X_i &= \sum_{k=1}^{K} z_{ik}\delta_{\omega_k} \\ z_{ik} &\sim Bernoulli(\pi_k) \\ \pi_k &\sim Beta(\frac{a}{K}, 1) \end{aligned} \tag{12}$$

Here we assume that each feature is independent from each other and could be selected by each data with same probability, i.e., $B_0(\mathcal{B}_k) = 1/K$ for $k = 1, 2, \ldots, K$ regions. Extending one parameter Beta distribution $Beta(\frac{a}{K}, 1)$ to two parameters Beta distribution $Beta(\frac{a}{K}, b\frac{K-1}{K})$, we may obtain more flexible model for our data.

Let $p(y_i|\mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta})$ denote a neighborhood-based classifier parameterized by $\boldsymbol{\theta}$, representing the probability of class label y_i for $\mathbf{x}_i$, given the neighborhood of $\mathbf{x}_i$ [10]. The semi-supervised classifier is defined as a mixture

$$p(y_i|\mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta}) = \textstyle\sum_{j=1}^{n} b_{ij}\, p^*(y_i|\mathbf{x}_j, \boldsymbol{\theta}) \tag{13}$$

Let $p^*(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ be a base classifier parameterized by $\boldsymbol{\theta}$, which gives the probability of class label y_i of data point $\mathbf{x}_i$. The base classifier can be implemented by any parameterized probabilistic classifier. For binary classification with $y \in \{1, 0\}$, the base classifier can be chosen as a probit regression

$$p^*(y_i = 1|\mathbf{x}_i, \boldsymbol{\theta}) = p(z_i >= 0|\mathbf{x}_i, \boldsymbol{\theta})$$

$$= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp(\frac{-||z_i - \boldsymbol{\theta}^T \mathbf{x}_i||^2}{2}) dz_i \quad (14)$$

where a constant element 1 is assumed to be prefixed to each $\mathbf{x}$(the prefixed $\mathbf{x}$ is still denoted as $\mathbf{x}$ for notational simplicity), and thus the first element in $\boldsymbol{\theta}_m$ is a bias term. b_{ij} represents the probability walking from $\mathbf{x}_i$ to $\mathbf{x}_j$ in t steps [10].

Let $\mathcal{L} \subseteq \{1, 2, \cdots, n\}$ denote the index set of labeled data in $\mathcal{X}$. The neighborhood-conditioned likelihood function is written as

$$\begin{aligned} & p\big(\{y_i, i \in \mathcal{L}\} | \{\mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\}, \boldsymbol{\theta}\big) \\ &= \prod_{i \in \mathcal{L}} p(y_i | \mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta}) = \prod_{i \in \mathcal{L}} \sum_{j=1}^{n} b_{ij}\, p^*(y_i | \mathbf{x}_j, \boldsymbol{\theta}) \end{aligned} \quad (15)$$

Suppose we are given M tasks, defined by M partially labeled data sets

$$\mathcal{D}_m = \{\mathbf{x}_i^m : i = 1, 2, \cdots, n_m\} \cup \{y_i^m : i \in \mathcal{L}_m\}$$

for $m = 1, \cdots, M$, where y_i^m is the class label of $\mathbf{x}_i^m$ and $\mathcal{L}_m \subset \{1, 2, \cdots, n_m\}$ is the index set of labeled data in task m. We consider M semi-supervised classifiers, parameterized by $\boldsymbol{\theta}_m$, $m = 1, \cdots, M$, with $\boldsymbol{\theta}_m$ responsible for task m.

Assuming that, given $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$, the class labels of different tasks are conditionally independent. Substituting (14) into (15) the joint likelihood function over all tasks can be written as

$$\begin{aligned} & p\big(\{y_i^m, i \in \mathcal{L}_m\}_{m=1}^M | \{\mathcal{N}_t(\mathbf{x}_i^m) : i \in \mathcal{L}_m\}_{m=1}^M, \{\boldsymbol{\theta}_m\}_{m=1}^M\big) \\ = & \prod_{m=1}^{M} \prod_{i \in \mathcal{L}_m} \sum_{j=1}^{n_m} b_{ij}^m\, p^*(y_i^m | \mathbf{x}_j^m, \boldsymbol{\theta}_m) \\ = & \prod_{m=1}^{M} \prod_{i \in \mathcal{L}_m} \sum_{j=1}^{n_m} b_{ij}^m \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp(\frac{-||z_{ij}^m - \boldsymbol{\theta}_m^T \mathbf{x}_j^m||^2}{2}) dz_{ij}^m \end{aligned} \quad (16)$$

where the m-th term in the product is taken from (15), with the superscript m indicating the task index. Note that the neighborhoods are built for each task independently of other tasks, thus a random walk is always restricted to the same task (the one where the starting data point belongs) and can never traverse multiple tasks.

Our objective is to learn $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$ jointly, sharing information between tasks as appropriate; at the same time, we also hope that within each cluster (a group of similar tasks), the classifier could automatically exclude irrelevant features and focus on useful ones.

Since $\sum_{j=1}^{n_m} b_{ij} = 1$, $\sum_{j=1}^{n_m} b_{ij}^m\, p^*(y_i^m | \mathbf{x}_j^m, \boldsymbol{\theta}_m)$ in (16) can be treated as a mixture model and we can get rid of the summation by introducing a hidden variables t_i^m – the index of datum to which $\mathbf{x}_i^m$ may transit. The generative model for the semi-supervised Multi-task feature learning with probit model can be written as

$$\begin{aligned} (y_i^m | t_i^m, z_{it_i^m}^m) &\sim I(z_{it_i^m}^m \geq 0)\delta_1 + I(z_{it_i^m}^m < 0)\delta_0 \\ (z_{it_i^m}^m | t_i^m, \mathbf{x}_{t_i^m}^m, \boldsymbol{\theta}_m) &\sim \mathcal{N}(\boldsymbol{\theta}_m^T \mathbf{x}_{t_i^m}^m, 1) \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\theta}_m &= \boldsymbol{\Theta}_m \circ W_m \\
\boldsymbol{\Theta}_m &\sim G \\
G &\sim \mathcal{DP}(\alpha_0 G_0) \\
G_0 &\sim BeP(B) \\
B &\sim \mathcal{BP}(a, b, B_0) \\
W_m &\sim \mathcal{N}(0, \beta_0 I)
\end{aligned} \tag{17}
$$

where $m = 1, 2, \ldots M, i = 1, 2, \ldots \mathcal{L}_m$ and the symbol $\circ$ represents the Hadamard, or elementwise multiplication of two vectors. In this generative model, we decompose classifier $\boldsymbol{\theta}_m$ into two parts, $\boldsymbol{\Theta}_m$ and W_m. A Dirichlet prior $\mathcal{DP}(\alpha_0 G_0)$ is imposed over $\boldsymbol{\Theta}_m$, which gives $\boldsymbol{\Theta}_m, m = 1, 2, \ldots, M$ clustering property based on the discussion above. The base measure G_0 is a draw from Beta process $\mathcal{BP}(a, b, B_0)$. G_0 is then a set of elements taking value $\{0, 1\}$. From the above background discussion, the proposed model has such nice properties: (1) Similar tasks will group together to share one $\boldsymbol{\Theta}^*_m$; Instead of using a pre-fixed number of clusters, the number of clusters is inferred from data itself. (2) Since G_0 is a draw from a Beta process, $\boldsymbol{\Theta}^*_m$ may be treated as a controller for feature learning, selecting relevant features and excluding irrelevant features. Data from tasks within one cluster will have the same feature selection mechanism, i.e., keeping or throwing away the same features across all tasks within one cluster. The tasks from different clusters then have different feature selection mechanism. (3)The introduction of parameters W_m gives the model more flexibility – it allows the different weights for those selected features for different tasks in one cluster.

Employing the stick-break construction for DP and truncating it to level N, as well as approximating Beta process with the finite latent feature models, the proposed model (17) can be written as

$$
\begin{aligned}
(y^m_i | t^m_i, z^m_{it^m_i}) &\sim I(z^m_{it^m_i} \geq 0)\delta_1 + I(z^m_{it^m_i} < 0)\delta_0 \\
(z^m_{it^m_i} | t^m_i, \mathbf{x}^m_{t^m_i}, \eta_m = h, \boldsymbol{\Theta}, \mathbf{W}_m) &\sim \mathcal{N}((\boldsymbol{\Theta}_h \circ \mathbf{W}_m)^T \mathbf{x}^m_{t^m_i}, 1) \\
(\eta_m | \mathbf{V}) &\sim \sum_{h=1}^{N} \pi_h \delta_h \\
\pi_h &= V_h \prod_{l<h} (1 - V_l), h = 1, 2, \ldots, N \\
(V_h | \alpha_0) &\sim Beta(1, \alpha_0) \\
(\alpha_0 | \tau_{10}, \tau_{20}) &\sim Ga(\tau_{10}, \tau_{20}) \\
(\boldsymbol{\Theta}_{hd} | \tau_d) &\sim Bernoulli(\tau_d) \\
(\tau_d | a, b) &\sim Beta(a, b) \\
(\mathbf{W}_m | \beta_0) &\sim \mathcal{N}(0, \beta_0 \mathbf{I})
\end{aligned} \tag{18}
$$

Here we impose a Gamma hyper-prior $Ga(\tau_{10}, \tau_{20})$ over concentration parameter α_0 for the DP prior. And W_m is independent drawn from a Gaussian distribution with mean zero and covariance matrix $\beta_0 \mathbf{I}$. The full likelihood function of the model is

$$
\begin{aligned}
& p(y, t, z, \eta, \boldsymbol{\Theta}, \mathbf{W}, \mathbf{V}, \boldsymbol{\tau}, \alpha_0, \alpha | \mathbf{x}, \tau_{10}, \tau_{20}, \zeta_{10}, \zeta_{20}, \beta_0) \\
&= \prod_{m=1}^{M} \Big(\prod_{i=1}^{\mathcal{L}_m} p(y^m_i | t^m_i, z^m_{it^m_i}) p(t^m_i, z^m_{it^m_i} | \mathbf{x}^m_{t^m_i}, \eta_m, \boldsymbol{\Theta}, \mathbf{W}_m) \Big) p(\eta_m | \mathbf{V}) p(\mathbf{W}_m | \beta_0)
\end{aligned}
$$

$$\prod_{h=1}^{N} p(\mathbf{V}_h|\alpha_0)\Big(\prod_{d=1}^{D} p(\mathbf{\Theta}_{hd}|\boldsymbol{\tau}_d)p(\boldsymbol{\tau}_d|a,b)\Big)p(\alpha_0|\tau_{10},\tau_{20}) \tag{19}$$

The sequential update equations of the Gibbs sampler are as follows.

- Draw z_{ij}^m from truncated normal distribution $\mathcal{TN}(z_{ij}^m|U_j^m,1,y_i^m z_{ij}^m)$, where $U_j^m = (\mathbf{\Theta}_{\eta(m)} \circ \mathbf{W}_m)^T \mathbf{x}_j^m$;
- Draw η_m from multinomial distribution with parameter π_m

$$\begin{aligned} \pi_{mh} &\propto \exp\left\{\sum_{i=1}^{\mathcal{L}_m}\sum_{j=1}^{n_m}\delta_{ij}^m\Big(z_{ij}^m\hat{\mathbf{W}}_h^T\mathbf{x}_j^m - \frac{1}{2}((\mathbf{x}_j^m)^T\hat{\mathbf{W}}_h\hat{\mathbf{W}}_h^T x_j^m)\Big)\right\} \\ &\times \exp\left\{\ln V_h + \sum_{l<h}\ln(1-V_l)\right\} \end{aligned} \tag{20}$$

  with $\hat{\mathbf{W}}_h = \mathbf{\Theta}_h \circ \mathbf{W}_m$ and $\pi_{mh} = \frac{\pi_{mh}}{\sum_{k=1}^N \pi_{mk}}$;
- Draw $\mathbf{W}_m$ from Gaussian Distribution with mean $\mu_{wm} = \Sigma_{wm}\Big(\sum_{i=1}^{\mathcal{L}_m}\sum_{j=1}^{n_m}\delta_{ij}^m z_{ij}^m(\hat{\mathbf{\Theta}}_{\eta(m)}\circ \mathbf{x}_j^m)\Big)$ and covariance $\Sigma_{wm} = \Big(\sum_{i=1}^{\mathcal{L}_m}\sum_{j=1}^{n_m}\delta_{ij}^m(\hat{\mathbf{\Theta}}_{\eta(m)}\circ\mathbf{x}_j^m)(\hat{\mathbf{\Theta}}_{\eta(m)}\circ\mathbf{x}_j^m)^T + \frac{I}{\beta_0}\Big)^{-1}$, where $\hat{\mathbf{\Theta}}_h = [\mathbf{\Theta}_h, \mathbf{\Theta}_h, \ldots, \mathbf{\Theta}_h]$;
- Draw V_h from Beta distribution $Beta(\upsilon_{h1},\upsilon_{h2})$ with $\upsilon_{h1} = 1 + \sum_{m=1}^M \mathbf{1}(\eta_m = h)$ and $\upsilon_{h2} = \alpha_0 + \sum_{m=1}^m \sum_{l>h}\mathbf{1}(\eta_m = l)$;
- Draw α_0 from gamma distribution $Gamma(\tau_1,\tau_2)$ with $\tau_1 = N - 1 + \tau_{10}$ and $\tau_2 = \tau_{20} - \sum_{h=1}^{N-1}\ln(1-\mathbf{V}_h)$;
- Draw $\mathbf{\Theta}_{hd}$ from Bernoulli distribution with parameter $p = \frac{1}{1+\exp(-tmp)}$, where $tmp = \sum_{m=1,\eta(m)=h}^M \sum_{i=1}^{\mathcal{L}_m}\sum_{j=1}^{n_m}\delta_{ij}^m\Big(z_{ij}^m\mathbf{W}_{md}\mathbf{x}_{jd}^m - \frac{1}{2}\mathbf{W}_{md}^2(\mathbf{x}_{jd}^m)^T\mathbf{x}_{jd}^m - \mathbf{\Theta_h}\circ\mathbf{W}_m^T\mathbf{x}_j^m\mathbf{W}_{md}\mathbf{x}_{jd}^m + \mathbf{W}_{md}^2\mathbf{\Theta}_{hd}\mathbf{x}_{jd}^{m\,2}\Big) + \ln(\tau_d) - \ln(1-\tau_d)$
- Draw τ_d from beta distribution $Beta(\tau_{d1},\tau_{d2})$ with $\tau_{d1} = a + \sum_{h=1}^N \mathbf{1}(\mathbf{\Theta}_{hd} = 1)$ and $\tau_{d2} = b + \sum_{h=0}^N \mathbf{1}(\mathbf{\Theta}_{hd} = 0)$;

In this subsection we derive the variational Bayesian approximation of the exact posterior distribution.

For simplicity the collection of all available data including features and labels is denoted as $\mathcal{D}$, the collection of all hidden variables and model parameters as $\mathbf{\Phi}$ and the collection of hyper-parameters as $\mathbf{\Psi}$. In our model $\mathcal{D} \equiv \{y, \mathbf{x}\}$, $\mathbf{\Phi} \equiv \{t, z, \eta, \mathbf{\Theta}, \mathbf{W}, \mathbf{V}, \boldsymbol{\tau}, \alpha_0\}$ and $\mathbf{\Psi} \equiv \{\tau_{10}, \tau_{20}, a, b, \beta_0\}$. By Bayes'law the joint posterior distribution of parameters $\mathbf{\Phi}$ given observed data $\mathcal{D}$ and hyper-parameters $\mathbf{\Psi}$ is

$$p(\mathbf{\Phi}|\mathcal{D},\mathbf{\Psi}) = \frac{p(\mathcal{D}|\mathbf{\Phi})p(\mathbf{\Phi}|\mathbf{\Psi})}{p(\mathcal{D}|\mathbf{\Psi})} \tag{21}$$

where $p(\mathcal{D}|\mathbf{\Psi}) = \int p(\mathcal{D}|\mathbf{\Phi})p(\mathbf{\Phi}|\mathbf{\Psi})d\mathbf{\Phi}$ often involves high-dimensional, complicated integrals. The variational Bayesian approach [11], [12], [13] approximate the joint posterior $p(\mathbf{\Phi}|\mathcal{D},\mathbf{\Psi})$ with a variational distribution $q(\mathbf{\Phi})$. The log of marginal likelihood

is written as

$$\begin{aligned}\ln p(\mathcal{D}|\mathbf{\Psi}) &= \ln \int p(\mathcal{D}|\mathbf{\Phi})p(\mathbf{\Phi}|\mathbf{\Psi})d\mathbf{\Phi} = \ln \int q(\mathbf{\Phi})\frac{p(\mathcal{D}|\mathbf{\Phi})p(\mathbf{\Phi}|\mathbf{\Psi})d\mathbf{\Phi}}{q(\mathbf{\Phi})} \\ &\geq \int q(\mathbf{\Phi}) \ln \frac{p(\mathcal{D}|\mathbf{\Phi})p(\mathbf{\Phi}|\mathbf{\Psi})d\mathbf{\Phi}}{q(\mathbf{\Phi})} = \mathcal{L}(\mathcal{D}|\mathbf{\Phi}) \end{aligned} \tag{22}$$

where $\mathcal{L}(\mathcal{D}|\mathbf{\Phi})$ is the low bound of $\ln p(\mathcal{D}|\mathbf{\Psi})$. The problem of computing posterior can be reformulated as an optimization problem of minimizing the Kullback-Leibler distance between $q(\mathbf{\Phi})$ and $p(\mathbf{\Phi}|\mathcal{D}, \mathbf{\Psi})$, which is equivalent to maximizing the lower bound $\mathcal{L}(\mathcal{D}|\mathbf{\Phi})$. The optimization problem can be analytically solved based on two assumptions on $q(\mathbf{\Phi})$ (i) $q(\mathbf{\Phi})$ is factorized; (ii) the integral over $\mathbf{\Phi}$ of $q(\mathbf{\Phi})$ should be equal to one. With appropriate choice of the form of the prior, the variational distribution of parameters $q(\mathbf{\Phi})$ are as follows.

- Update $q(t_i^m = j, z^m_{it_i^m})$ for $m = 1, 2, \ldots, M, i = 1, 2, \ldots, \mathcal{L}_m, j = 1, 2, \ldots, n_m$:

$$q(t_i^m = j, z^m_{it_i^m}) \propto b_{ij}^m \mathcal{TN}(z_{ij}^m | U_j^m, 1, y_i^m z_{ij}^m) \tag{23}$$

  where $U_j^m = \sum_{h=1}^N \rho_{mh}(\langle \mathbf{\Theta}_h \rangle \circ \langle \mathbf{W}_m \rangle)^T \mathbf{x}_j^m$;
- Update $q(t_i^m = j) = \delta_{ij}^m = \frac{p(y_i^m | x_j^m, U_j^m) b_{ij}^m}{\sum_{k=1}^{n_m} p(y_i^m | x_j^m, U_j^m) b_{ik}^m}$
- Update $q(\eta_m = h) = \rho_{mh}$, where

$$\begin{aligned}\rho_{mh} \propto \quad & \exp\left\{ \sum_{i=1}^{\mathcal{L}_m} \sum_{j=1}^{n_m} \delta_{ij}^m \left(\langle z_{ij}^m \rangle \langle \hat{\mathbf{W}}_h \rangle^T \mathbf{x}_j^m - \frac{1}{2}((\mathbf{x}_j^m)^T \langle \hat{\mathbf{W}}_h \hat{\mathbf{W}}_h^T \rangle x_j^m) \right) \right\} \\ & \times \exp\left\{ \langle \ln V_h \rangle + \sum_{l<h} \langle \ln(1 - V_l) \rangle \right\} \end{aligned} \tag{24}$$

  with $\hat{\mathbf{W}}_h = \mathbf{\Theta}_h \circ \mathbf{W}_m$, $\langle \hat{\mathbf{W}}_h \rangle = \langle \mathbf{\Theta}_h \rangle \circ \langle \mathbf{W}_m \rangle$ and $\langle \hat{\mathbf{W}}_h \hat{\mathbf{W}}_h^T \rangle = \langle (\mathbf{\Theta}_h \mathbf{\Theta}_h^T) \circ (\mathbf{W}_m \mathbf{W}_m^T) \rangle$. After Normalizing, we obtain $\rho_{mh} = \frac{\rho_{mh}}{\sum_{k=1}^N \rho_{mk}}$;
- Update $q(\mathbf{W}_m)$; The variational posterior of $\mathbf{W}_m$ can be shown to be normal with covariance $\Sigma_{wm} = \left(\sum_{i=1}^{\mathcal{L}_m} \sum_{j=1}^{n_m} \delta_{ij}^m \sum_{h=1}^N \frac{\rho_{mh}}{n_m} (\langle \hat{\mathbf{\Theta}}_h \hat{\mathbf{\Theta}}_h^T \rangle \circ (\mathbf{x}_j^m \mathbf{x}_j^{mT})) + \frac{I}{\beta_0} \right)^{-1}$ and mean $\mu_{wm} = \Sigma_{wm} \left(\sum_{i=1}^{\mathcal{L}_m} \sum_{j=1}^{n_m} \delta_{ij}^m \sum_{h=1}^N \rho_{mh} \langle z_{ij}^m \rangle (\langle \hat{\mathbf{\Theta}}_h \rangle \circ \mathbf{x}_j^m \right)$, where $\hat{\mathbf{\Theta}}_h = [\mathbf{\Theta}_h, \mathbf{\Theta}_h, \ldots, \mathbf{\Theta}_h]$ and

$$\langle \hat{\mathbf{\Theta}}_h \hat{\mathbf{\Theta}}_h^T \rangle = n_m \begin{bmatrix} \theta_{11}^2 & \theta_{11}\theta_{12} & \ldots & \theta_{11}\theta_{1D} \\ \theta_{11}\theta_{12} & \theta_{12}^2 & \ldots & \theta_{12}\theta_{1D} \\ \ldots & & & \\ \theta_{11}\theta_{1D} & \theta_{12}\theta_{1d} & \ldots & \theta_{1D}^2 \end{bmatrix}$$

- Update $q(\mathbf{V}_h)$; the variational posterior of $\mathbf{V}_h$ is also a Beta distribution $Beta(\upsilon_{h1}, \upsilon_{h2})$ with $\upsilon_{h1} = 1 + \sum_{m=1}^M \rho_{mh}$ and $\upsilon_{h2} = \langle \alpha_0 \rangle + \sum_{m=1}^m \sum_{l>h} \rho_{mh}$.
- Update $q(\alpha_0)$; the variational posterior of α_0 can be shown to be a Gamma distribution $Gamma(\tau_1, \tau_2)$ with $\tau_1 = N - 1 + \tau_{10}$ and $\tau_2 = \tau_{20} - \sum_{h=1}^{N-1} \langle \ln(1 - \mathbf{V}_h) \rangle$
- Update $q(\mathbf{\Theta}_{hd})$; the probability of $\mathbf{\Theta}_{hd}$ equal to 1 is proportional to

$$q(\mathbf{\Theta}_{hd} = 1)$$

$$\propto \exp\left\{\sum_{m=1}^{M}\sum_{i=1}^{\mathcal{L}_m}\sum_{j=1}^{n_m}\delta_{ij}^m\rho_{mh}\Big(\langle z_{ij}^m\rangle\langle \mathbf{W}_{md}\rangle\mathbf{x}_{jd}^m-\frac{1}{2}\langle\mathbf{W}_{md}^2\rangle(\mathbf{x}_{jd}^m)^T\mathbf{x}_{jd}^m\Big)\right\}$$
$$\times\exp\left\{\sum_{m=1}^{M}\sum_{i=1}^{\mathcal{L}_m}\sum_{j=1}^{n_m}\delta_{ij}^m\rho_{mh}\Big(\mathbf{W}_{md}^2\boldsymbol{\Theta}_{hd}\mathbf{x}_{jd}^{m\,2}-\langle\boldsymbol{\Theta}_{\boldsymbol{h}}\circ\mathbf{W}_m\rangle^T\mathbf{x}_j^m\mathbf{W}_{md}\mathbf{x}_{jd}^m\Big)\right\}$$
$$\times\quad\exp\left\{\langle\ln(\tau_d)\rangle\right\} \tag{25}$$

and the probability of $\boldsymbol{\Theta}_{hd}$ equal to 0 is proportional to $\exp\left\{\langle(1-\ln(\tau_d))\rangle\right\}$;

- Update $q(\tau_d), d=1,2,\ldots,D$; the variational posterior of τ_d is still a Beta distribution $Beta(\tau_{d1},\tau_{d2})$ with $\tau_{d1}=\sum_{h=1}^{N}q(\boldsymbol{\Theta}_{hd}=1)+a$ and $\tau_{d2}=\sum_{h=0}^{N}q(\boldsymbol{\Theta}_{hd}=0)+b$.

We here also discuss how active learning may be incorporated into this framework. We take an information-theoretic approach to identifying the data locations at which the labels would be most informative to the classifier parameters. Our approach is based on use of Fisher information [14], [15], which is related to previous uses of active learning [16], [17] as applied to purely supervised models. The Fisher information involves the log-likelihood; as a result the prior is excluded from the calculation. Since the tasks are connected through the prior, this implies that calculation of Fisher information can be performed for each individual task separately (*not* independently though, since the true parameters are replaced by their most recent estimates, as seen below, which are coupled by the prior). Therefore, we drop each variable's independence on task index m, for notational simplicity. The data log-likelihood is obtained by taking the logarithm of (15),

$$\begin{aligned}\ell(\boldsymbol{\theta}) &\stackrel{Def.}{=} \ln p\big(\{y_i, i\in\mathcal{L}\}|\{\mathcal{N}_t(\mathbf{x}_i):i\in\mathcal{L}\},\boldsymbol{\theta}\big)\\ &=\sum_{i\in\mathcal{L}}\ln\sum_{j=1}^{n}b_{ij}\,p^*(y_i|\mathbf{x}_j,\boldsymbol{\theta})\end{aligned} \tag{26}$$

where the base classifier is assumed as above to be a logistic-regression classifier, *i.e.*, $p^*(y_i|\mathbf{x}_j,\boldsymbol{\theta})=[1+\exp\{-g(\mathbf{x}_i,\boldsymbol{\theta})\}]^{-1}$. By definition [15], the Fisher information matrix (FIM) for the data likelihood is

$$\begin{aligned}&FIM\left\{p\big(\{y_i,i\in\mathcal{L}\}|\{\mathcal{N}_t(\mathbf{x}_i):i\in\mathcal{L}\},\boldsymbol{\theta}\big)\right\}\\ &=\mathbb{E}_{\{y_i\}_{i\in\mathcal{L}}}\left[\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]\left[\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]^T\\ &=\sum_{i\in\mathcal{L}}\mathbb{E}_{y_i}\left[\frac{\sum_{j=1}^{n}b_{ij}p^*(y_j=y_i|\mathbf{x}_j,\boldsymbol{\theta})p^*(y_j=-y_i|\mathbf{x}_j,\boldsymbol{\theta})y_i\mathbf{x}_j}{\sum_{k=1}^{n}b_{ik}p^*(y_k=y_i|\mathbf{x}_k,\boldsymbol{\theta})}\right]\\ &\quad\times\left[\frac{\sum_{j=1}^{n}b_{ij}p^*(y_j=y_i|\mathbf{x}_j,\boldsymbol{\theta})p^*(y_j=-y_i|\mathbf{x}_j,\boldsymbol{\theta})y_i\mathbf{x}_j}{\sum_{k=1}^{n}b_{ik}p^*(y_k=y_i|\mathbf{x}_k,\boldsymbol{\theta})}\right]^T\\ &=\sum_{i\in\mathcal{L}}\frac{\mathbf{z}_i\mathbf{z}_i^T}{\gamma_i}\end{aligned} \tag{27}$$

where

$$\mathbf{z}_i\stackrel{Def.}{=}\sum_{j=1}^{n}b_{ij}p^*(y_j=1|\mathbf{x}_j,\boldsymbol{\theta})p^*(y_j=-1|\mathbf{x}_j,\boldsymbol{\theta})\mathbf{x}_j \tag{28}$$

$$\gamma_i \overset{Def.}{=} \sum_{k=1}^{n} b_{ik} p^*(y_j = 1|\mathbf{x}_k, \boldsymbol{\theta}) \sum_{k=1}^{n} b_{ik} p^*(y_j = -1|\mathbf{x}_k, \boldsymbol{\theta}) \tag{29}$$

Assume that $\mathbf{x}_* \in \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \setminus \mathcal{L}$ is a newly labeled sample and added to the labeled set, so $\mathcal{L}$ changes to $\widetilde{\mathcal{L}}$. The Fisher information matrix changes to

$$\begin{aligned} & FIM\left\{p\big(\{y_i, i \in \widetilde{\mathcal{L}}\}|\{\mathcal{N}_t(\mathbf{x}_i) : i \in \widetilde{\mathcal{L}}\}, \boldsymbol{\theta}\big)\right\} \\ &= \sum_{i \in \mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i} + \frac{\mathbf{z}_* \mathbf{z}_*^T}{\gamma_*} \end{aligned} \tag{30}$$

The ratio of the determinants (one of several possible measures [14], related to the entropy under a Gaussian approximation for the posterior for the model parameters) of the old and new FIMs is

$$\begin{aligned} & \frac{\det\left[\sum_{i \in \mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i} + \frac{\mathbf{z}_* \mathbf{z}_*^T}{\gamma_*}\right]}{\det\left[\sum_{i \in \mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i}\right]} \\ &= \left(1 + \frac{1}{\gamma_*} \mathbf{z}_*^T \left[\sum_{i \in \mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i}\right]^{-1} \mathbf{z}_*\right) \det\left[\sum_{i \in \mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i}\right] \end{aligned} \tag{31}$$

The logarithmic difference is

$$\psi(\mathbf{x}_*) = 1 + \frac{1}{\gamma_*} \mathbf{z}_*^T \left[\sum_{i \in \mathcal{L}} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\gamma_i}\right]^{-1} \mathbf{z}_* \tag{32}$$

which we employ as our selection criterion in identifying the most informative data sample for labeling. The criterion $\psi(\mathbf{x}_j)$ is calculated for all $\mathbf{x}_j \in \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \setminus \mathcal{L}$, and the one with the maximum is the most informative data location to obtain a label. The true value of $\boldsymbol{\theta}$ required in calculating $\mathbf{z}$ and γ is replaced with the most recent update of the parameters, following the strategy taken in [17], [14]. To the best of our knowledge, this is the first use of active learning in an MTL setting, and we are also considering a semi-supervised model.

RESULTS

To evaluate the proposed multi-task feature learning algorithm, experimental results are presented on four data sets, one based on synthetic data and others on benchmark real data. In order to compare with other algorithms, we employ AUC as the performance measure, where AUC stands for area under the receiver operation curve (ROC)[18]. The relation the AUC and the error rate is discussed in [19].

Throughout this section, the basic setup prior hyper-parameters are as follows: $\beta_0 = 1$, $\tau_{10} = 0.05$, $\tau_{20} = 0.05$, $a = 1$, $b = 1$ and W_m are set according to sample means. The truncation level for Dirichlet process is the number of tasks and the initial number of latent features are the dimension of features.

We first demonstrate the proposed multi-task feature learning model on a synthetic data, for illustrative purpose. For our synthetic example, we have six tasks $\{\mathbf{x}^m, y^m\}, m =$

$1, 2, \ldots, 6$. The features $\mathbf{x}^m$ for each task are generated from a Gaussian distribution with mean zeros and covariance identity matrix. The dimension of datum is 50 and we generated $N = 500$ samples for each task. Assume the label $y^m(i), i = 1, 2, \ldots, 500$ is generated from $y^m(i) = sign(\boldsymbol{\theta}_m^T \mathbf{x}^m(i) + \mathcal{N})$. $\mathcal{N}$ is an additive white Gaussian noise with a signal-to-noise ratio of 10. For tasks 1, 2 and 3 feature index $\{3, 7, 12, 14, 16, 31, 33, 47, 48, 50\}$ are relevant for classification, which means that linear classifiers $\boldsymbol{\theta}_m, m = 1, 2, 3$ are very sparse with only elements $\{3, 7, 12, 14, 16, 31, 33, 47, 48, 50\}$ non-zero. For tasks 4, 5 and 6 feature index $\{11, 14, 22, 28, 30, 32, 33, 38, 39, 40\}$ are useful for obtaining the correct classifiers. The truncation level of Dirichlet process for the synthesized data is equal to six. Since ground truth is available for this synthesized example, it is employed as a starting point for analysis of the models.

The ground truth of classifier coefficients $\boldsymbol{\theta}_m, m = 1, 2 \ldots 6$ as well as classifier coefficients learnt with the proposed semi-supervised feature learning algorithm are depicted in Figure 1. In the Gibbs sampling implementation, the burning period is 1000 and the results are the average of 500 iterations after burning period. The results of variational Bayesian is the average of 100 random trials.

From Figure 1, we can see that the proposed semi-supervised MTL feature learning algorithm can correctly select useful features for each task and can also obtain a very good approximation for those weights.

Each curve in Figure 2(a) represents the mean AUC as a function of the number of labeled data. Here we compare our proposed algorithm to the Semi-supervised MTL [10]. Figure 2(b) gives us the sharing patterns that semi-supervised MTL feature learning algorithm finds for the six tasks. we plot the Hinton diagram [20] of the between-task sharing matrix found by the semi-supervised MTL. The (i, j)-th element of sharing matrix records the number of occurrences that task i and j are grouped into the same cluster. The Hinton diagram in Figure 2(b) also shows the agreement of the sharing mechanism of the semi-supervised MTL with the similarity between the tasks.

As seen from Figure 1 and Figure 2, three observations are made:

- With the feature selection mechanism, the proposed algorithm may help to improve classification performance;
- Task 1, 2 and 3 are grouped together and task 4, 5 and 6 are grouped together, which is in the agreement of the ground truth;
- The number and positions of zero-elements for task 1, 2 and 3 are almost same, although the non-zero elements have different values; The number and positions of zero-elements for task 4, 5 and 6 are almost same, although the non-zero elements have different values;

In the second example we consider the underwater-mine classification problem studied in [21], where the acoustic imagery data were collected with four different imaging sonars from two different environments (see [21] for details)[1]. This is a binary classification problem aiming to separate mines from non-mines based on the synthetic-aperture sonar (SAS) imagery. For each sonar image, a detector automatically finds the objects of interest, and a 13-dimensional feature vector is extracted for each target. The number

[1]The data for the underwater mine example are available at www.ece.duke.edu/ lcarin/UnderwaterMines.zip

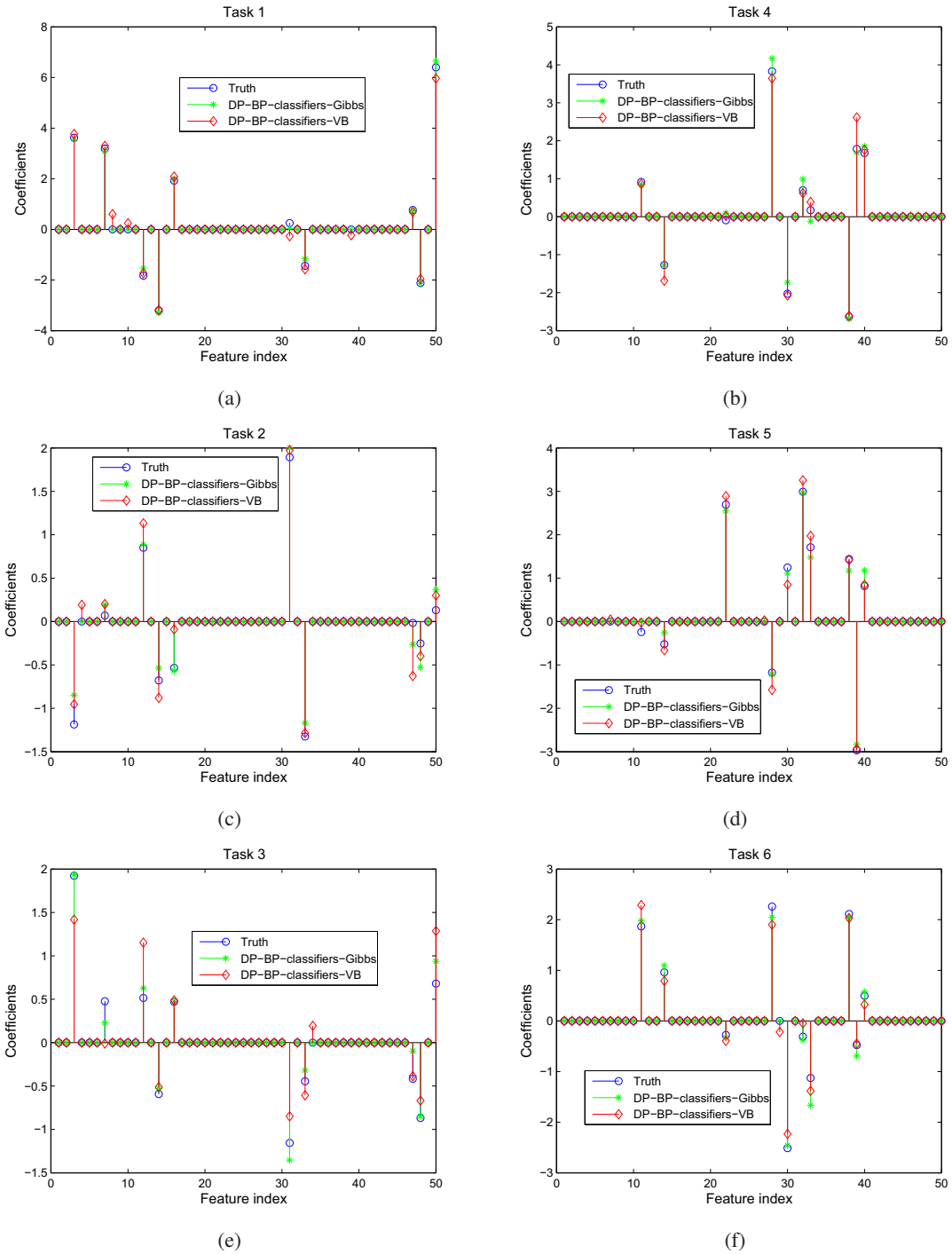
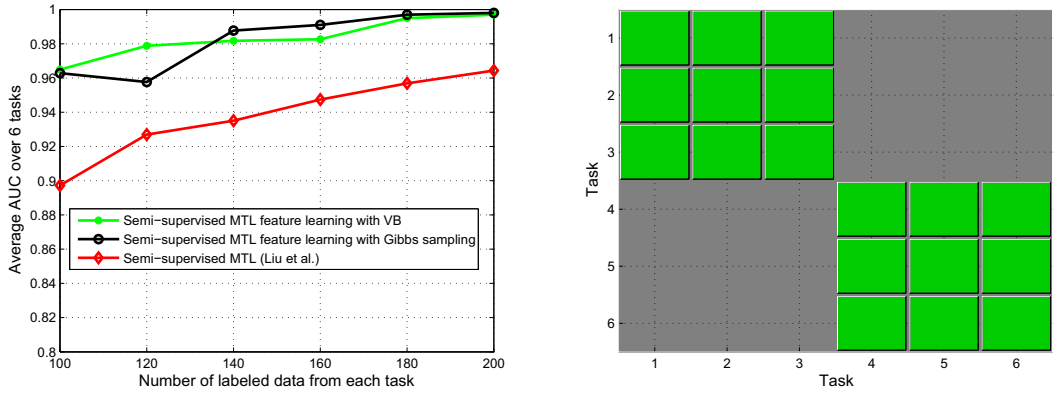


Fig. 1. Classifier coefficients of the semi-supervised MTL feature learning algorithm on Tasks 1-6. The horizontal axis is feature index in each task. The vertical axis is classifier coefficients. The coefficients for synthesized data are donated by blue circles; the learnt coefficients using Gibbs sampler are donated with green stars; the learnt coefficients with variational Bayesian (VB) are donated with red diamond.

of mines in each of the eight tasks varies from 9 to 65, and each task contains from 10 to 100 times more non-mines (clutter) than mines.

In this problem, there are a total of 8 data sets, constituting 8 tasks. The 8 data sets



(a) (b)

Fig. 2. (a) Performance of the semi-supervised MTL feature learning algorithm on Tasks 1-6. The horizontal axis is the number of labeled data in each task. The vertical axis is the AUC averaged over the six tasks;(b) The Hinton diagram of between-task sharing found by semi-supervised MTL feature learning.

are collected with four sonar sensors from two different environmental conditions. The total number of data points in each task is listed in Table I. The distribution of sensors

TABLE I

NUMBER OF DATA POINTS IN EACH TASK FOR THE UNDERWATER-MINE DATA SET CONSIDERED IN FIGURE **??**.

Task ID	1	2	3	4	5	6	7	8
Number of data	1813	3562	1312	1499	2853	1162	1134	756

and environments are listed in Table II.

It can be seen from the synthesized example that variational Bayesian (VB) approximation can achieve almost the same performance as Gibbs sampler but with much higher speed. In this subsection we only present results of VB approach. To show the strength of the proposed algorithm, we add 7 dummy features to the original feature vector and extend a 13-dimensional feature vector to a 20-dimensional feature vector.

Following [21], we perform 50 independent trials, in each of which we randomly select a subset of data for which labels are assumed known (labeled data), train the semi-supervised MTL with feature selection and without feature selection [10], and test the classifiers on the remaining data. The AUC averaged over 8 tasks is presented in Figure 3, as a function of the number of labeled data in each task, where each curve represents the mean calculated from the 50 independent trials. Figure 4 gives us one sample of the weights of 8 tasks when labeled data from each task is 30.

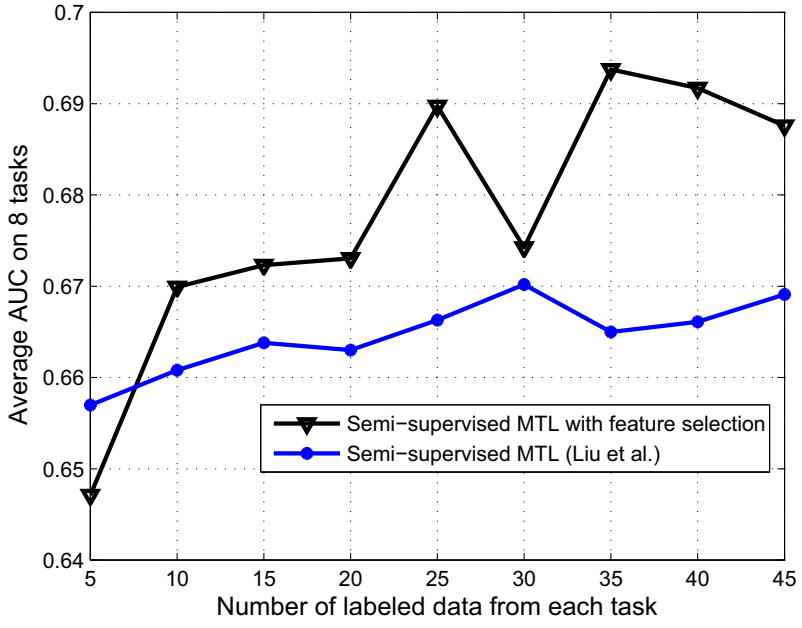
The results on underwater target classification also show that the proposed semi-supervised MTL feature learning outperforms the semi-supervised MTL algorithm [10] by selecting relevant features.

To demonstrate active learning integrated with multi-task semi-supervised learning,

TABLE II

DISTRIBUTION OF SENSORS AND ENVIRONMENTS OVER 8 TASKS. ENVIRONMENT A IS RELATIVELY CHALLENGING WHILE ENVIRONMENT B RELATIVELY BENIGN, WITH THESE CHARACTERISTICS MANIFESTED BY THE DETAILS OF THE SEA BOTTOM.

Task	Sonar	Environment
1	1	B
2	1	A
3	2	B
4	3	B
5	4	B
6	2	A
7	3	A
8	4	A

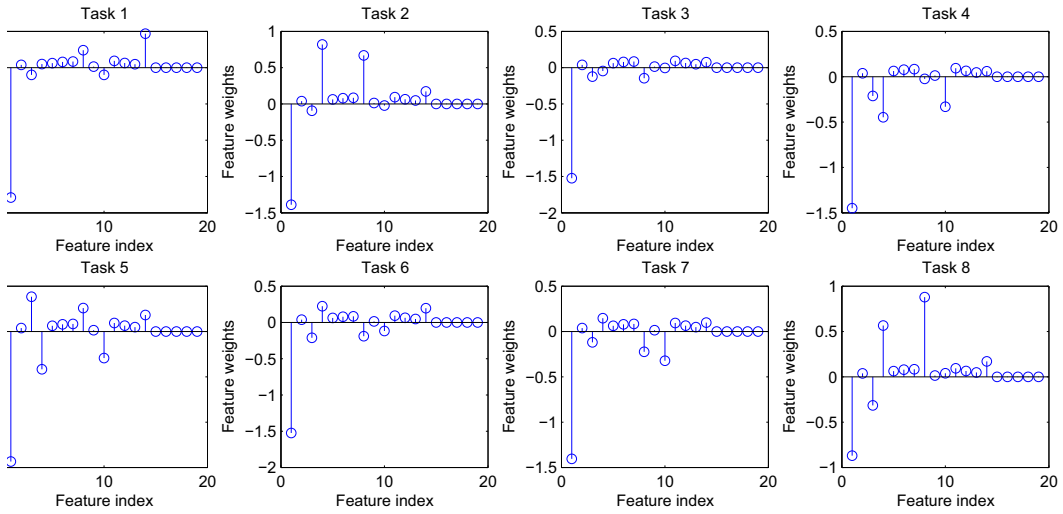


(a)

Fig. 3. Performance of the semi-supervised MTL feature learning algorithm on underwater target classification, in comparison to semi-supervised MTL algorithm [10]. The horizontal axis is the number of labeled data in each task. The vertical axis is the AUC averaged over the six tasks and 50 independent trials.

we consider a remote sensing problem based on data collected from real landmine fields[2]. In this problem there are a total of 19 sets of data, collected from various landmine fields (with inert landmine simulants). Each data point is represented by a 9-dimensional feature vector extracted from synthetic aperture radar images. Since this is a detection problem, the class labels are binary, with 1 indicating landmine and 0 indicating clutter (false alarm). We have also demonstrated this technology with MCM

[2]The data from the landmine example are available at www.ece.duke.edu/ lcarin/LandmineData.zip



(a)

Fig. 4. One sample of weights when labeled data from each task is 30.

data, as discussed above.

As opposed to the setting in [10], where it is assumed that labeled data from the 19 data sets are available simultaneously, we here assume the much more realistic case for which labeled data are acquired sequentially within one data set (task) at one time. Once the process of label acquisition in a given environment is completed, that environment is not revisited to acquire new labeled data.

Each of the 19 data sets defines a task, in which we aim to find landmines with a minimum number of false alarms. Of the 19 data sets, 1-10 are collected at foliated regions and 11-19 are collected at regions that are bare earth or desert. We expect fewer new labeled data when considering a new task for which environmental conditions stay unchanged from previous tasks (but this is *inferred* by the algorithm, and not imposed by the user).

In the experiment both labeled and unlabeled data are used in training the algorithm. After training, the algorithm is tested on the unlabeled data to calculate the area under ROC curve (AUC) for each data set. We compare the active-learning results with AUC results obtained using random selection of labeled data. For the case where the labeled data are randomly selected, we perform 20 independent trials, and compute the mean as well as error bars of AUC from the trials. Since the data sets are acquired sequentially, the results are presented as AUC as a function of the number of tasks from which labeled data are acquired (the ordering of the tasks is arbitrary; the task order considered here was selected as to make a point on the number of labels actively acquired, as discussed further below).

We observe from the results in Figure 5 that active learning performs much better than random selection for a small number of data sets (tasks). As discussed below, the total number of labels used in random selection of labels is the same as that used for active learning. When the number of tasks increases, the benefit of active learning diminishes since the scarcity of labeled data is overcome via multi-task learning.

In Figure 6 we plot the number of labeled data for each task, as a function of task index. For the active-learning algorithm the total number of labeled data is $n = 174$, across all 19 tasks (this is determined adaptively, by the proposed algorithm). For the

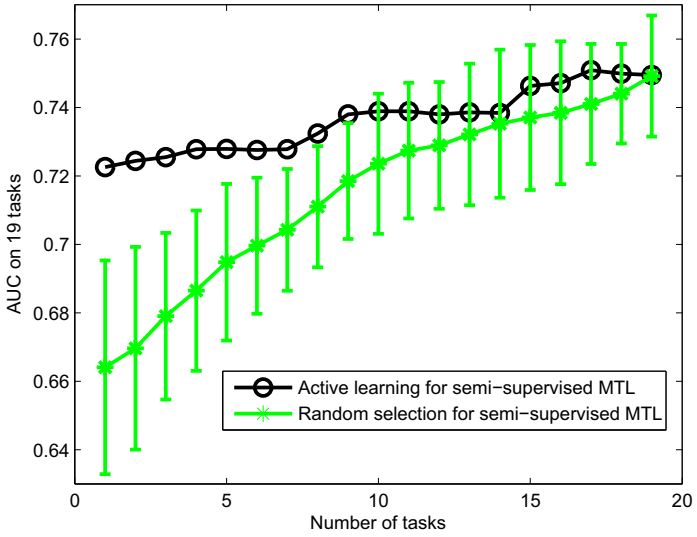


Fig. 5. Performance of active learning for semi-supervised MTL algorithm in comparison to semi-supervised MTL with randomly-selected labeled data. The horizontal axis is the number of tasks from which labeled data are acquired. The vertical axis is the AUC averaged over the 19 tasks.

random selection of labeled data, the data from all 19 tasks are put together, and 174 samples are selected at random for labeling; therefore, the number of labels acquired per task is not constant (the data in Figure 6, for random selection, represents one example). For the active-learning results in Figure 6, note the big jump in the number of labeled data at task $k = 11$. Recall from above that data sets 1-10 are from generally foliated regions and data sets 11-19 are from regions that are generally bare earth or desert. Therefore, the jump in Figure 6 at $k = 11$ is consistent with expectations based on the properties of the environments.

IMPACT/APPLICATIONS

The technology is of significant utility for MCM and ASW applications. We have developed a new classification algorithm for multi-task feature learning. By utilizing the clustering property of Dirichlet process (DP) and feature selection property of Beta process, our algorithm can learning classifiers jointly, sharing information as appropriate; at same time, the algorithm can automatically exclude irrelevant features and learn good weights for relevant features for each task.

TRANSITIONS

The technology is being transitioned from SIG to the Navy, via a collaboration with Dr. Robert McDonald, of NSWC Panama City.

RELATED PROJECTS

SIG is executing a related SBIR on active learning.

PUBLICATIONS

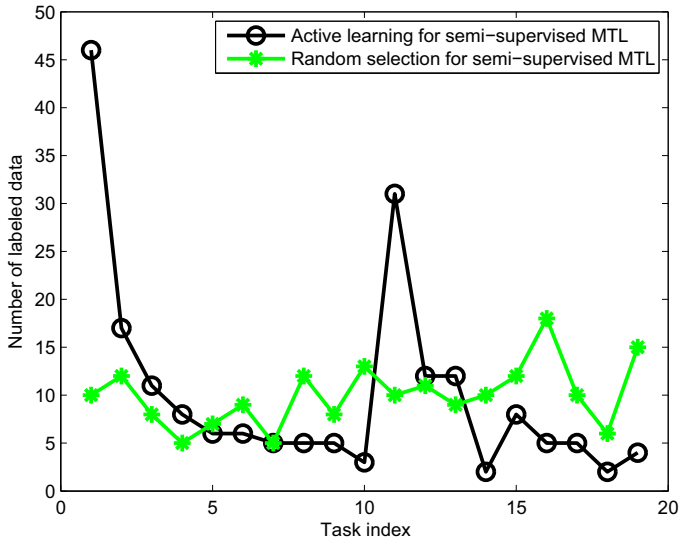


Fig. 6. Number of labeled data using active learning in comparison to number of labeled data with random selection; for the latter, this is one random example.

Q. Liu, X. Liao, H. Li, J. Stack and L. Carin, "Semi-supervised multitask learning," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 31, pp. 1074-1086, June 2009.

J. R. Stack, G. Dobeck, Xuejun Liao, L. Carin, "Kernel-Matching Pursuits With Arbitrary Loss Functions", *IEEE Transactions on Neural Networks*, Vol 20, No 3, pp. 395-405, March 2009

H. Li, X. Liao and L. Carin, "Active learning for semi-supervised multi-task learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2009.

Also had a JUA paper published in 2009, details omitted.

PATENTS

None.

HONORS

None.

REFERENCES

[1] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[2] D. Blackwell and J. MacQueen, "Ferguson distributions via polya urn schemes," *Annals of Statistics*, vol. 1, pp. 353–355, 1973.

[3] C.E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.

[4] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, 1994.

[5] H. Ishwaran and L.F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, pp. 161–173, 2001.

[6] N. L. Hjort, "Nonparametric Bayes estimators based on Beta processes in models for life history data," *Annals of Statistics*, vol. 18, no. 3, pp. 1259–1294, 1990.

[7] J. Paisley and L. Carin, "Nonparametric factor analysis with Beta process priors," in *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[8] R. Thibaux and M. I. Jordan, "Hierachical Beta process and the Indian buffet process," in *International Conference on Artificial Intelligence and Statistics*, 2007.

[9] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Idian buffet process," in *Advances in Neural Information Processing System*, 2005.

[10] Q. Li, X. Liao, and L. Carin, "Semi-supervised multi-task learning," in *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2009, MIT Press.

[11] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M.I. Jordan, Ed. MIT Press, Cambridge, 1999.

[12] T.S. Jaakkola and M.I. Jordan, "A variational approach to Bayesian logistic regression models and their extensions," in *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

[13] Z. Ghahramani and M. Beal, "Propagation algorithms for variational Bayesian learning," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. MIT Press, Cambridge, MA, 2001.

[14] V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, New York, 1972.

[15] T.M. Cover and J.A. Thomas, *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.

[16] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[17] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 589–603, 1992.

[18] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[19] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Scholkopf, Eds. 2004, vol. 16, MIT Press, Cambridge, MA.

[20] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in boltzmann machines," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, Eds. 1986, vol. 1, pp. 282–317, MIT Press, Cambridge, MA.

[21] J. R. Stack, F. Crosby, R. J. McDonald, Y. Xue, and L. Carin, "Multi-task learning for underwater object classification," in *Proceedings of the SPIE defense and security symposium*, 2007, vol. 6553, pp. 1–10.